IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

Adaptive Robot-Human Handovers with Preference Learning

Gojko Perovic *,1, Francesco Iori *,1,2, Angela Mazzeo[2], Marco Controzzi†,2, Egidio FaloticFo†,1

***Abstract*—This paper proposes an adaptive method for robot-to-human handovers under different scenarios. The method combines Dynamic Movement Primitives (DMP) with Preference Learning (PL) to generate online trajectories that are reactive to human motion, modulating the speed of the robot. The PL allows for tuning the coupling parameters of the DMP, tailoring the interaction to each participant personally, and allowing for qualitative analysis of user preferences. Simulation of an interaction-constrained learning task with different optimization techniques is performed to determine an appropriate learning approach for a handover task. The validity of the approach is demonstrated through experiments with participants on two handover tasks, with results indicating that the proposed method leads to seamless and pleasurable interactions.**

***Index Terms*—Machine Learning for Robot Control; Human-Robot Collaboration; Human-Aware Motion Planning**

I. INTRODUCTION

HANDOVER is a fundamental part of most of the common cooperative tasks. Thus, humans perform it intuitively, exchanging the objects with ease, and coordinating both spatially and temporally. As such, emulating smooth and pleasurable handovers in Human Robotic Cooperation (HRC) tasks is imperative for successful interactions.

People's prior knowledge of handovers enables them to anticipate each other's movements and adjust accordingly [1]. While both spatial and temporal precision are integral for successful applications of the robotic handover, humans tend to place more importance on temporal precision [2]. Temporal coordination tends to be more challenging in practical HRC settings where the handover could be susceptible to interruptions or perturbations due to changing dynamics in human motion (Fig. 1). These perturbations could arise from scene misinterpretation on the side of the high-level controller (e.g. the robot initiating a handover while the human is not ready to receive the object), unexpected disturbances on the human side, or simply user disengaging from the handover. Furthermore, users might have different preferences in how the robotic partner should move based on different settings or personal comfort with the robot. Thus, it can be beneficial to develop a system that allows participants to tune the behavior of the robot intuitively.

To this end, we propose an adaptive method based on Dynamic Movement Primitives (DMP) [3] with Preference Learning (PL) for dynamic handovers. DMP allow for online trajectory generation which is reactive to human motion, modulating the speed of the robot [4]. Further, PL is used to tune the parameters of the DMP (and thus the adaptive capabilities of the robot) from interactive user feedback. In doing so, generated trajectories can be more coordinated, responsive, and robust to perturbations, thus ensuring seamless and pleasurable interaction. The proposed approach enables tailoring the interaction to each participant personally, and thus allows for qualitative analysis of user preferences between changing handover scenarios.

II. RELATED WORK

A. Trajectory Generation for Handovers

Handover represents a joint action between two partners, cooperating both spatially and temporally [5]. As such, to accommodate for changes in the environment and partner behavior, adaptable behavior is crucial. Pre-planned methods may only be effective if all the constraints are known, which is not the case in the present study. To this end, several approaches have been used in the literature, including DMP [6], [7], Probabilistic Movement Primitives [8], and Interaction Primitives [9]. DMP are a popular choice for trajectory generation in robotics due to their ability to generate smooth and continuous trajectories while being able to handle perturbations and noise.

In our previous work [4], we focused on a DMP-based approach that coupled the evolution of the robot trajectory to the human hand trajectory. In that work [4], and present work as well, we consider adapting to unmodeled perturbations *given the permanence of the handover intention* on the giver's side. However, in our previous study, we hand-picked the relevant DMP parameters, which may not be optimal for all users or situations. In this work, we aim to enable humans to optimize toward preferred behavior, thus resulting in more natural and intuitive handovers, across two different handover scenarios.

B. Coordination, Perturbations, and Preferences in Handovers

Koene et al. [2] have demonstrated that temporal precision is a major determinant of user perception of the handover.

Manuscript received: March 13, 2023; Revised: May 29, 2023; Accepted: July 24, 2023. This paper was recommended for publication by Editor Aleksandra Faust upon evaluation of the Associate Editor and Reviewers' comments.

*These authors contributed equally to this work. †MC and EF also contributed equally to this work

[1]Supported by the European Commission under the Horizon 2020 framework program for Research and Innovation under the Specific Grant Agreement No. 945539 (Human Brain Project SGA3).

[2]Supported by the European Commission under the Horizon 2020 framework program for Research and Innovation (project acronym: APRIL, project number: 870142).

All authors are with The BioRobotics Institute and Department of Excellence in Robotics, Scuola Superiore Sant'Anna, Pisa, Italy.

Digital Object Identifier (DOI): see top of this page.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

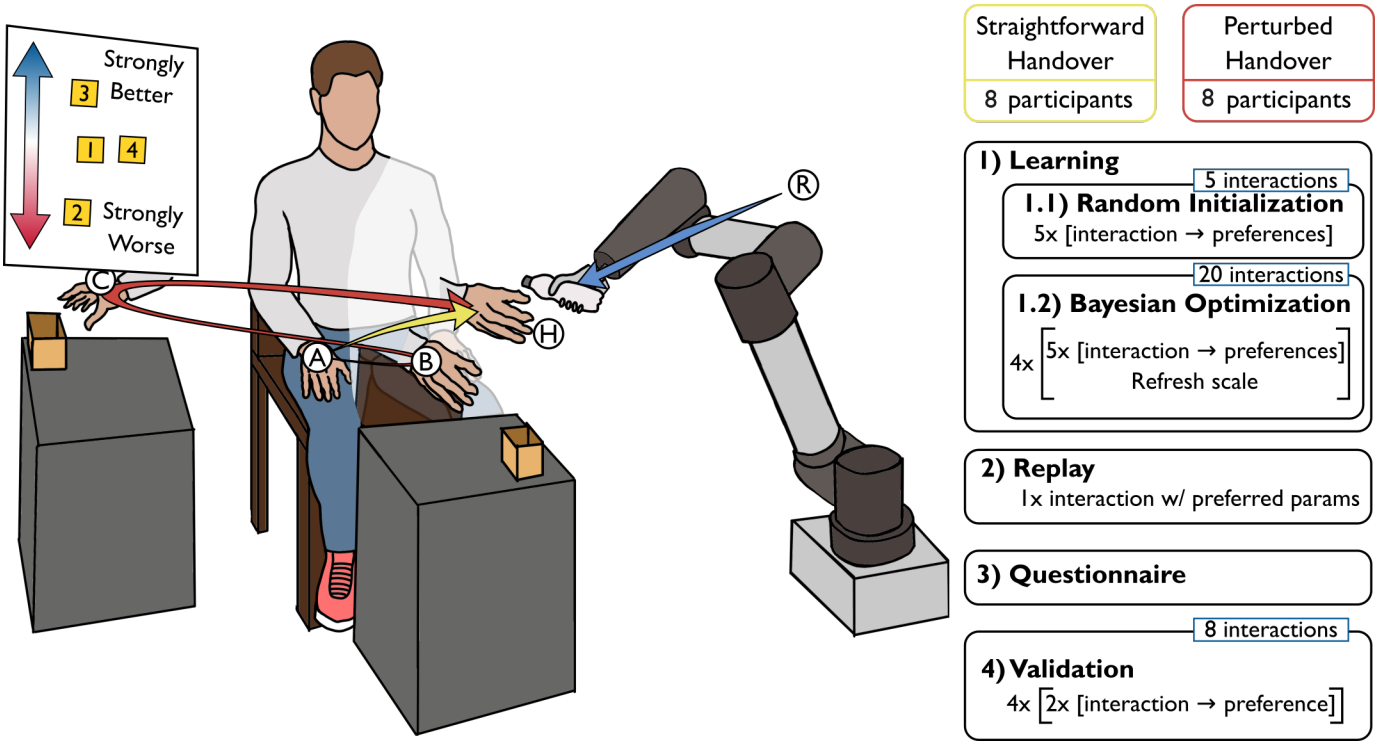

Fig. 1. Experimental setup (left) and protocol (right) for two different handover scenarios. In Straightforward Handover (denoted by the yellow trajectory), the user, starting from the start position (A), is tasked to reach the handover exchange location (H). In the Perturbed Handover (denoted by the red trajectory), the user, starting from the start position (A), is tasked with taking the box at location (B), transferring it to location (C), and then reaching the handover exchange location (H). Robot always starts in the robot start location (R) and reaches (H). 16 participants are divided into two groups of 8 and follow the outlined protocol. The board with a relative scale is placed behind the participants to give their preferences after the relevant interactions. In the displayed example, upon obtaining the preference on interaction labeled as "4", comparisons added to $\mathcal{D}$ are $\theta^4 \succ \theta^2$ and $\theta^3 \succ \theta^4$. As interactions "1" and "4" are rated as equal, a comparison is not constructed between them.

Considering the handover with a predetermined exchange location allows the controller to more easily adapt to changes in the speed of the human hand, thus better examining the effects of temporal coordination between the human and the robot.

While the literature on robotic handover seldom considers interactions in which the handover action could be disrupted by another task or unexpected perturbations, Huang et al. [10] propose an adaptive controller that takes into account the availability of the human receiver. The aforementioned approach is based on a finite state machine (FSM) which consists of multiple robot states and as such address a similar control problem on a higher level.

Using human preferences to directly shape the generation of the trajectory is often limited by the high dimensionality of the problem and the difficulty to gather significant feedback. In [11] the authors propose a method to learn user preferences from real-world interaction using the contextual policy search. However, this method relies on absolute feedback, which might introduce the problems of drift, change of scale, or forget [12], [13]. Handover tasks could also successfully be learned from demonstrations as shown by Wu et al. [7]. Similarly, they employ DMP to generate the trajectory. The focus of the work in [7] is on the spatial coordination between partners, and in human-to-robot handovers. In both of these works [11], [7], handovers are represented as straightforward interactions that are not susceptible to possible perturbations.

To address the issues introduced by absolute feedback, the preferred behavior can also be learned from user preferences expressed as binary comparisons [12], [13]. A notable line of work on pairwise comparisons applied to robotic learning is demonstrated in research by Tucker et al. [14], [15]. In this work, comparisons are supplemented by coactive feedback to improve the sample efficiency of the model optimizing the gait of exoskeletons and bipedal robots [14], [15]. Such improvements could benefit the proposed method in the future as well.

It is worth noting that a more elaborate method for employing preference feedback could be set up by learning a reward which some reinforcement-based learning algorithm could exploit [16], [17]. However, in this case, additional challenges could be introduced, both computationally and in terms of the necessary training data. Furthermore, [18] have shown that DMP can be adapted by learning on discrete human feedback. As the handover is relatively simple in terms of task complexity, we aim to use preference learning to adapt the dynamics of the system by changing critical parameters of the trajectory generation module.

III. DMP COORDINATION WITH PL

A. Dynamic Movement Primitives

DMP can be used to generate trajectories online as an evolution of a virtual dynamical system. The original framework

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

[3] is commonly extended to learn and reproduce both periodic and non-periodic trajectories.

A one-dimensional trajectory can be represented with one degree of freedom (DOF) $x(t)$, with initial state $x(t_0) = x_0$ and desired final goal $x(t_f) = g$. The DMP models the evolution of this trajectory as a second-order dynamical system:

$$\tau^2 \ddot{x} = \alpha_x(\beta_x(g - x) - \tau\dot{x}) + f(s) \tag{1}$$

where α_x and β_x are positive parameters, and τ a time constant. With $\beta_x = \frac{\alpha_x}{4}$, a critically damped system is obtained. $f(s)$ represents a forcing term that can be used to mold the evolution of the trajectory and is defined as:

$$f(s) = f_{nl} s(g - x_0) \tag{2}$$

where f_{nl} is a arbitrary non-linear function approximator and $(g - x_0)$ is a scaling factor. As the system evolves, a phase variable s decreases monotonically from 1 to 0, and $x(t)$ converges stably toward g, as the effect of the forcing term vanishes. The phase s typically follows a first-order dynamics of the form:

$$\tau\dot{s} = -\alpha_s s \tag{3}$$

with τ and α_s again as the time-constant of the system and a positive parameter, respectively. The system can be extended to multiple DOFs by defining one transformation system for each DOF and a shared canonical system. By doing so, different DOFs are coordinated by the single common phase variable.

B. Coupling Terms

By adding spatial or temporal *coupling terms* DMP framework can be extended to produce elaborate behavior. The coupling terms are integrated as:

$$\tau^2 \ddot{x} = \alpha_x(\beta_x(g - x) - \tau\dot{x}) + f(s) + C_s \tag{4}$$

$$\tau = \tau_0(1 + C_t) \tag{5}$$

where C_s represents the spatial and C_t the temporal coupling term.

To coordinate the trajectory with the behavior of the user, this works uses the method proposed in [4].

Given an estimate of the handover location g, d is defined as the distance of the human hand, with an initial value $d(t_0) = d_0$ at the start of the interaction. A second-order low-pass filter is applied on the measured distance d (denoted as $\tilde{d}$):

$$d = \|g - x_{hand}\|_2 \tag{6}$$

$$\ddot{\tilde{d}} = \alpha_d \left(\frac{\alpha_d}{4}(d - \tilde{d}) - \dot{\tilde{d}} \right) \tag{7}$$

Coupling terms are then defined to be:

$$C_t = k_t \sigma_d \left(\frac{\tilde{d}}{d_0} \right) \sigma_{\dot{d}} \left(\dot{\tilde{d}} \right) \tag{8}$$

$$C_s = -\alpha_x \tau\, \dot{x}\, k_s \sigma_d \left(\frac{\tilde{d}}{d_0} \right) \sigma_{\dot{d}} \left(\dot{\tilde{d}} \right) \tag{9}$$

with

$$\sigma_i(y) = \frac{1}{1 + \exp\left(-a_i\left(y + \delta_i\right)\right)} \tag{10}$$

where k_t and k_s are positive gains and $\sigma_i(y)$ is a sigmoid function with x-axis offset δ_i and steepness coefficient a_i. The sigmoid $\sigma_{\dot{d}}$ is mainly responsible for the adaptation to human motion, while σ_d reduces its influence once the hand reaches the final position. As the main role of σ_d is as a filter against practical edge cases [4], here the values of a_d and δ_d are arbitrarily fixed to 13.0 and -0.35, respectively, producing a quasi-linear response for $\left(\frac{\tilde{d}}{d_0}\right) \in [0.2, 0.5]$, as in [4].

C. Preferential Learning with Refreshing Scale

Learning from human feedback is difficult due to inherent problems in human evaluations. Absolute human feedback is usually noisy and unreliable, suffering from *drift* (scale shifting over time) and *anchoring* (early interactions being deemed as more important) [12], [13]. Moreover, different users can have very different internal scales. To address these issues, participants can be asked to give a relative evaluation, stating how the most recent interaction compared to the previous one [12], [13]. To this end, a probit model can be used to infer the utility function of human preferences u from binary feedback [12], [13]. In the proposed approach, we assume that u is a latent value given by the user's perception of the interaction relative to the changing robot trajectory. Given a data set of ranked pairs:

$$\mathcal{D} = \{\theta_i^r \succ \theta_i^c; i = 1, ..., m\} \tag{11}$$

where $\theta_i^r, \theta_i^c \in \Theta$ are instances of points in the parameter space. After collecting the data, a zero-mean non-parametric Gaussian process (GP) prior can be fitted as:

$$\mathcal{P}(\mathbf{u}) = |2\pi\mathbf{K}|^{\frac{1}{2}} \exp\left(\frac{1}{2}\mathbf{u}^T\mathbf{K}^{-1}\mathbf{u}\right) \tag{12}$$

where $\mathbf{u} = |u(\theta_1), u(\theta_2), ..., u(\theta_n)|^T$ is the utility of user choice at sampled points and $\mathbf{K}$ is the $n \times n$ covariance matrix (n is the number of instances) [12]. To estimate the posterior distribution of u given $\mathcal{D}$, model is fit [12], [13]:

$$\mathcal{P}(\mathbf{u}|\mathcal{D}) \propto \mathcal{P}(\mathbf{u}) \prod_{i=1}^{m} \mathcal{P}(r_i > c_i | u(r_i), u(c_i)) \tag{13}$$

This problem can be treated as an optimization of an expensive-to-estimate black-box function, a setting where Bayesian Optimization (BO) can be employed effectively [13]. Most commonly in BO settings, Expected Improvement (EI) acquisition function can be used to efficiently sample the next set of parameters. Given the best observed value of the latent function $u(\theta^*)$, new point is queried by maximizing the EI:

$$\theta = \arg\max_{\theta} \mathbb{E}(\max\{0, u_{t+1}(\theta) - u(\theta^*)\}|\mathcal{D}) \tag{14}$$

However, the performance of BO methods can be susceptible to noise, especially since human evaluations represent high noise feedback. A method proposed by Letham et al. [19] for batch Monte Carlo approximation of Expected Improvement under Noisy observations and constraints (qNEI) can alleviate

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

these hindrances. While EI-based acquisition functions are often regarded as greedy, in HRC tasks this property can be somewhat desired, as collecting data tends to be expensive, and the number of interactions is limited. The exploration-exploitation trade-off in an optimization problem with a limited number of interactions and scaling noise is thus considered in Sect. IV.

While the preferential feedback alleviates variance introduced by noise in human evaluation, preferences themselves are not highly informative. In a classic setting [13], two points are sampled and displayed, and the user gives a preference between them. Similarly, in virtual environments, it is possible to simultaneously provide a gallery of options from which the user selects the preferred option [13]. However, this is not possible in practical, HRC scenarios due to its interactive nature. To alleviate these disadvantages we propose a relative scale with a periodic refresh. Interactions are performed in batches of q sampled points ($q = 5$ in the experiments with participants outlined in Sect. V). The relative scale is presented as an arrow ranging from "Strongly Worse" to "Strongly Better". This design decision comes after the pilot study [20], where some participants found difficulties in expressing their preferences on a seven-point scale (e.g. asking to rate a later interaction as in-between two previous interactions that are already placed in consecutive bins). This representation is employed to reduce the strain on participants when giving the preferences. Participants are explained that it is only important to order the preferences relative to each other and that visual representation and the distance between the placed preferences carry no absolute value and do not affect the model.

After each interaction, pairwise comparisons are extracted between the most recent set of parameters and previous sets within the batch (Fig. 1), added to data $\mathcal{D}$, a new prior is obtained, and a new parameter point θ is sampled. Thus, comparisons are generated between up to q number of points, instead of a two-by-two approach, increasing the information gained from each interaction. After each batch the scale is refreshed, removing the user's feedback so far. Then, the user is presented with the previous best-observed point θ^*, and a new batch of sampled interactions begins. The benefits of the refresh are twofold: first, by refreshing hindrances of drift and anchoring are removed, as the scale does not have an absolute value; secondly, the strain on participants' memory is reduced, and they can focus on the relation between the few most recent interactions.

To this end, the proposed method aims to learn the parameters which shape the reactiveness of the generated trajectory: $a_{\dot{d}}$, $\delta_{\dot{d}}$, and $k_t = k_s = k$. This approach is slightly different when compared to the pilot study [20], where the temporal and spatial gains were decoupled. However, this could cause certain instabilities in the virtual dynamical system in cases where the difference between gains was significant, while for close enough values the differences in trajectory were not noticeable. The continuous parameter space Θ of the PL algorithm was set to $a_{\dot{d}} \in [1, 10.0]$, $\delta_{\dot{d}} \in [-1.0, 1.0]$, and $k \in [0.01, 15]$, with $\theta = (a_{\dot{d}}, \delta_{\dot{d}}, k)$. This again comes as a result of findings in the pilot study [20] where the original bounds were too broad, and certain parameter settings could lead to unsafe behavior (i.e. robot excessively accelerating resulting in high inertia). The values of the DMP are reported in Table I.

TABLE I
DMP'S PARAMETERS

Optimized (Θ)		
$a_{\dot{d}}$	$\delta_{\dot{d}}$	k
$[1, 10.0]$	$[-1.0, 1.0]$	$[0.01, 15]$

Fixed						
τ_0	α_x	α_s	α_d	a_d	δ_d	k_{ext}
1.0	20.0	4.0	20.0	13.0	-0.35	0.8

The complete block diagram of the proposed system is demonstrated in Fig. 2.

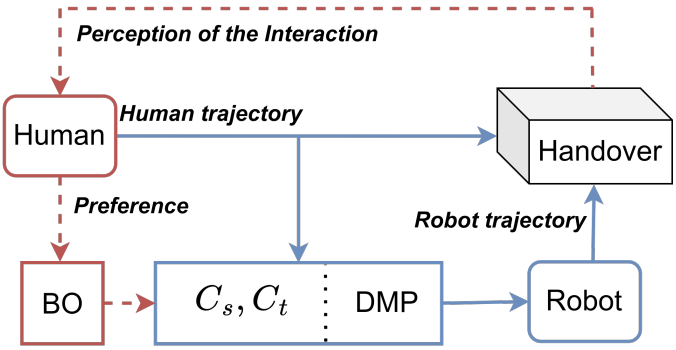

Fig. 2. Trajectory generation and learning loop diagram. The blue lines signify the trajectory generation and control loops, that act continuously during the handover. The red lines highlight the preferential learning method, which acts asynchronously after each interaction to sample the next combinations of parameters to try.

IV. SIMULATION OF AN INTERACTION CONSTRAINED LEARNING TASK

As it would be exceedingly time-consuming to examine the performance of the BO method under different hyperparameter settings in a real-world robotic scenario, a simulation is performed. The task is set up as reaching the randomly sampled point (θ^t) in a three-dimensional space by evaluating the l_2 norm:

$$u(\theta) = \|(\theta^t - \theta)\|_2 \tag{15}$$

By collecting observations as:

$$y(\theta) = u(\theta) + j \cdot \mathcal{N}(0, \sigma^2) \tag{16}$$

where $\mathcal{N}(0, \sigma^2)$ is Gaussian noise and j is the index of the interaction within the batch ($j = 1, ..., q$). Thus, comparisons are reconstructed as $\theta^r \succ \theta^c$ if $y(\theta^r) < y(\theta^c)$. By doing so, scaling noisy observations are simulated to represent possible participant forget as more interactions are performed within the batch.

Two of the most critical hyperparameters for the proposed PL loop are the acquisition function and the size of the batch q. To better evaluate the performance of the BO approach

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

under different acquisition functions, an additional acquisition function is considered. Given two queries, Expected Utility of Best Option (EUBO) aims to increase the information by maximizing the utility obtained due to these queries [21]. Thus, two acquisition functions, qNEI and EUBO are evaluated. The influence of batch size q is evaluated by considering qNEI approaches with $q = 2$ and $q = 5$. These conditions are tested in the noiseless ($\sigma = 0$) and noisy ($\sigma = 0.05$) settings. GP is implemented with a Radial Basis Function kernel (as in [12], [13], [22]).

The model is initialized by sampling 5 random parameter points. To simulate the constraints of real-world interactions, we only consider 25 unique queries in total from each approach. It is worth noting that practically in the two-by-two experiment ($q = 2$) this would result in 45 interactions, while with $q = 5$ it would result in 30 interactions. The experiment is run for a total of 500 trials for each test condition. To evaluate the performance and sample efficiency of different BO approaches the mean distance to the target by interaction and the success rate ($u(\theta) < 0.1$) by interaction are reported in Fig. 3.

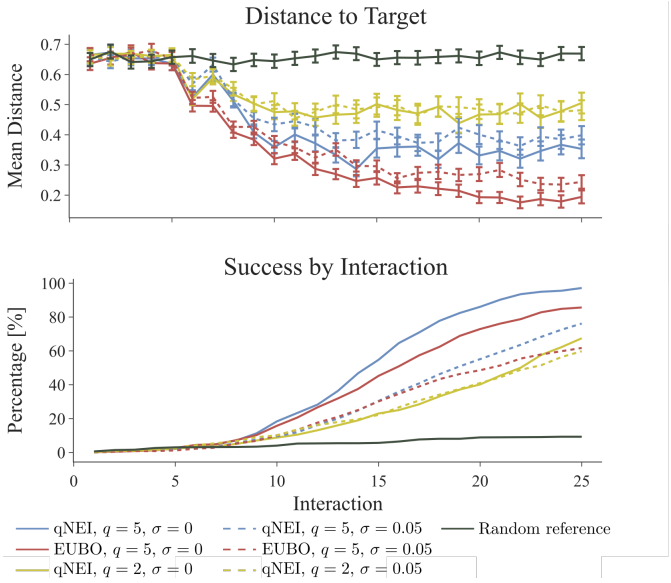

Fig. 3. Top: Mean distance to the target by interaction in the simulated task. Vertical lines represent the 95% confidence interval. Bottom: Success rate by interaction in the simulated task.

From the top plot in Fig. 3 it is apparent that the EUBO is a more elaborate acquisition function as the utility function is better approximated. However, in an HRC scenario, we might be interested in finding some set of parameters that leads to satisfactory performance in the limited interaction time users might have with the robot. Thus, from the bottom plot in Fig. 3 the merit of using qNEI in a practical scenario, with a reduced number of trials can be seen. Due to the "greediness" of qNEI, satisfactory solutions could be more easily reached. It is worth noting that this is highly dependent on the scenario, but it can be a reasonable approximation in HRC settings where the goal is to have a pleasurable interaction. Further, it is worth noting the performance of batched approach ($q = 5$), even in the presence of scaling noise. While the noise in evaluations is both dependent on the task and the user, the amount of information gained, and thus a potential reduction in real-world training time, could not be overlooked. Therefore it might be beneficial to perform interactions in refreshing batches, should the users experience no difficulty remembering the interactions.

V. EXPERIMENTS WITH PARTICIPANTS

In Section III we presented a method to optimize towards a preferred set of adaptive parameters for robot trajectories from human preferences in a limited number of handover interactions. An experiment is run to first learn the preferred set of parameters from each participant's feedback, collected over a limited number of interactions. Secondly, we evaluate the hypothesis that the participant prefers the set of parameters obtained by BO over a set of parameters picked with quasi-random Sobol sampling. Hence, the controlled variable is the set of parameters selected with PL, and the dependent variable is the percentage of BO sets evaluated as preferred over the sampled ones.

A. Handover Scenarios

The *straightforward handover* (yellow arrow in Fig. 1) consists of the user reaching for and grasping the object (an empty 0.5l water bottle) from the robot giver. The robot and the participant are given the "Go" signal simultaneously. Furthermore, they are positioned so that both the robot and the human perform their respective trajectories at the approximately same time, should the robot move at top speed. This can be considered as a standard handover, commonly discussed in the literature, without any perturbations.

In a second scenario (red arrow in Fig. 1), the *perturbed handover*, the participant is tasked with performing a secondary task before engaging in the handover. Again, both the robot and participant are given the "Go" signal simultaneously. In this scenario, the participant has to first reach for a box placed in the vicinity ($\sim$ 25cm) of the handover location. Secondly, they have to place the box on the desk nearby. Finally, after placing the box, the participant should reach for the final location to receive the object from the robotic giver, and place it in the aforementioned box.

B. Setup

To facilitate the robotic handover, Universal Robots UR5 CB-series manipulator equipped with IH2 Azzura hand (Prensilia SRL) [23] is used. A HEX-70-XE OnRobot six-axis force-torque sensor is mounted between the wrist and the hand, to enable a force-threshold release of the object. A Vicon motion capture system with 6 Bonita cameras (Vicon Ltd) is used to track the human hand at 100Hz. A speaker was used to give the participants an auditory "Go" signal. PL algorithm was implemented using BoTorch [22], and run on a separate machine.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

C. Experiment Protocol

Sixteen participants (right-handed, 8 female and 8 male, aged 23-40) took part in the experiment. Eight participants were assigned to each scenario. Informed consent for voluntary participation was obtained in accordance with the Declaration of Helsinki.

Firstly, motion capture markers were fit to the participant's hand. Then, the assigned task was explained. It was also explained to the participants that the robot has different coordination capabilities which might vary from trial to trial. Participants were asked to give their relative preference after each interaction. The choices for relative preference being rating interactions as better, worse, or equal to the previous within the batch, with an "equal" evaluation meaning that the participant does not have a clear preference between those evaluated as such. Fig. 1 reports an example of how pairwise comparisons are constructed after each preference. It was made clear that the robot behavior might change as a result of these preferences and how the relative scale functions. Further, it was emphasized to the participants that the release of the object was always going to be the same and that they should only give preference in relation to the timing of the robot's trajectory.

To start the experiment, 5 interactions are performed with randomly sampled parameters to initialize the model. Then, 4 learning batches of 5 interactions each are performed, with the scale refreshing after each interaction. Finally, 1 more interaction is performed with the best set of parameters learned, and a subjective questionnaire is given to the participants:

- Q1: It was easy to remember the interactions within the batch. (5 points from "Not at all" to "Perfectly")
- Q2: Did you feel that the robot was constantly improving? (Yes / No / Maybe)
- Q3: Relative to the last set of parameters tried, how much did you perceive that the robot was coordinating with you? (5 points from "Not at all" to "Perfectly")

Then, participants were asked to perform 4 more batches of 2 trials each. These served as validation trials, where unbeknownst to participants, the learned set of parameters was compared to randomly sampled parameters, in a randomly sampled order. Participants again had the choice to rate each interaction as better, worse, or equal to the previous one. During these validation trials, human and robot trajectories were recorded for qualitative analysis. In total, the experimental procedure lasted approximately 40 minutes per participant. This study was approved by the local ethical committee of the Scuola Superiore Sant'Anna, Pisa, Italy (approval number 21/2022).

VI. RESULTS

A. Validation results

The results of validation trials are presented in Table II for straightforward handover and in Table III for the perturbed scenario. The percentages of preferences toward the set of parameters selected with BO were significantly above the reference threshold of 50% (Sign test, N=16, $p< 0.001$). The reference threshold to test against was chosen as the median value of the possible outcomes.

TABLE II
VALIDATION RESULTS IN THE STRAIGHTFORWARD HANDOVER.

Participant	$a_{\dot{d}}$	$\delta_{\dot{d}}$	k	**Validation (%)**
1	1.00	0.29	0.01	**100**
2	10.00	-0.80	0.01	**75**
3	6.06	-1.00	0.01	***75**
4	1.00	0.40	0.01	**100**
5	6.82	-0.52	0.01	**75**
6	10.00	-1.00	0.01	**75**
7	4.81	-0.48	15	**100**
8	10	-0.33	15	**100**

TABLE III
VALIDATION RESULTS IN THE PERTURBED HANDOVER.

Participant	$a_{\dot{d}}$	$\delta_{\dot{d}}$	k	**Validation (%)**
9	1.00	-0.20	0.01	**50**
10	10.00	-0.51	12.83	**100**
11	4.19	-1	0.01	**100**
12	10.00	0.04	15	**100**
13	8.50	0.17	4.16	**75**
14	6.77	0.55	2.52	**100**
15	4.18	-0.11	9.93	**75**
16	7.49	0.43	15	**100**

B. Survey results

In Table IV results of subjective surveys in straightforward and perturbed handover are given on the questionnaire described in Sect. V-C. Answers to Q1 were above the threshold of 3 (Sign test, $p < 0.001$). This result points to the relative ease in remembering the interactions within the batch, and in general, participants did not claim any difficulties. Moreover, there is no significant difference between the answers in straightforward and perturbed handover (Mann-Whitney U Test), suggesting different tasks did not affect the ability to remember the interactions in the batch. No significant difference between the straightforward and perturbed handover was found for Q3 as well (Mann-Whitney U Test), underlying that a similar level of coordination was perceived in the two conditions. Q2 was posed as a precaution, as the trend in participants answering "Yes" to this question could mean that there are some biases towards learning robots, which could hinder the PL process. No statistical analysis is performed on this question as it would require a significantly higher number of respondents.

TABLE IV
SURVEY RESULTS

Straightforward				Perturbed			
Participant	Q1	Q2	Q3	Participant	Q1	Q2	Q3
1	4	Maybe	3	9	3	No	4
2	4	Maybe	4	10	3	Maybe	4
3	4	Yes	4	11	4	Yes	5
4	3	Yes	4	12	4	Yes	4
5	5	Maybe	5	13	3	No	4
6	4	Maybe	4	14	4	Maybe	5
7	3	Maybe	4	15	5	Maybe	5
8	4	No	5	16	3	No	4

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

C. Qualitative Metrics

To qualitatively assess the preferred robot behavior, human and robot trajectories are recorded. While it is challenging to define a direct optimization metric when unmodeled perturbations might arise, comparing the respective trajectories given the preferred parameters allows for a qualitative assessment of controller performance. The trajectory plots are represented in Fig. 4.

VII. DISCUSSION

Two handover scenarios are set up to investigate the correlation and contrast in participant preferences in different handover tasks. The motivation behind the task set-up in perturbed handover is not only to present a realistic, less structured, interaction but also to present a worst-case scenario for the controller. By placing the secondary object (the box) close to the final handover location, higher precision is required from the controller in interpreting human motion. Thus, learning the parameters is made more difficult as we hypothesize that users might prefer more nuanced reactive behavior in such a scenario. This is largely due to the fact that many parameter settings lead to non-reactive robot behavior.

By considering validation results in Sec. VI-A, it can be seen that the proposed PL approach has a high success rate. Out of all the trials, the only optimization which could be deemed as suboptimal is the one performed by Participant 9. In this case, the PL algorithm likely over-fitted to a local minimum representing a slightly-damped non-reactive controller, when from validation trials it was made clear that the participant might have preferred a more reactive behavior. This might be due to the correlation between speed and coordination, as even if the robot was not reactive it was still performing the trajectory at the right speed to be perceived as coordinated. Nevertheless, this exemplifies the difficulty of learning from direct human feedback, as participants might have to weigh between many different correlated characteristics when giving their preferences. Furthermore, it is worth noting that the challenges that come with noise in human evaluation, affect the validation process as well. For example, in Table II, in the single missed validation trial marked with *, the participant was presented with two (identical) non-reactive controllers, but preferred one to another, rather than rating them as the same. These findings might indicate that the assumptions from Sect. IV are valid, as it is very challenging to estimate a ”global” optimum of the latent utility function of the human preference. For the same reason, and due to limited experimental time with human participants, it is challenging to estimate the absolute number of interactions required for finding the optimal parameters. For the proposed scenarios and the required parameter space, 25 interactions could be viewed as a conservative estimate to reach satisfactory performance. Should the dimension of the required parameter space be significantly higher, the noisiness of evaluations would likely diminish the effectiveness of the BO approach. Considering all the factors, the PL approach seems to overcome these challenges in the proposed setting and consistently produce satisfactory controllers in the limited amount of interactions.

The questionnaire was given to participants to better understand their subjective experience. Mainly, it was of interest to verify that participants in general did not have trouble remembering the interactions within the batch. The handover does not represent a long task, so batching a small number of interactions can lead to more effective data usage, as the participants did not exhibit difficulty remembering the trials within the batch of five interactions. Q2 was posed to investigate if there was some bias toward a learning robotic agent, i.e. participants rating later interactions more favorably due to expectation of a robot that is constantly improving. This might represent a challenge as BO tends to query points with high variance, leading to points of varying perceived value. As mentioned, the number of participants does not allow for statistical analysis of this bias. However, further studies into biases introduced by interaction with learning robots are warranted, as this might improve the design of the learning methods. From Q3, no statistically significant difference can be reported in the perception of robot coordination between changing scenarios, even though the dynamics of the two tasks are varying. However, somewhat lower scores in answering this question might be due to a misunderstanding, as some participants later noted that they believed that a ”set of parameters” refers to the last learning batch and the converged parameters together, instead of converged parameters separately (as intended).

From qualitative metrics, there is an insight into pleasurable robot behavior between scenarios. In the straightforward scenario, 6 out of 8 participants converged to a completely non-reactive set of parameters of the robot moving at the top speed (leading to identical robot trajectories). Two remaining participants converged to highly reactive parameters, with high accelerations. From Fig. 4 in Participant 7 and Participant 8 plots it can be observed that the robot initiates its movement after the human. In the perturbed scenario, reactive behavior is more appreciated, as 5 out of 8 participants converged to reactive controllers. Between the participants, there were varying types of reactive parameters, for example, Participant 14 preferred a slightly-damped reactive controller resulting in smooth accelerations. On the other hand, Participant 16 preferred a highly reactive controller, with the robot moving only after the participant has placed the box on the desk (Fig. 1). It is worth again mentioning Participant 7 as they might have not converged to their global optimum, but instead would have preferred a reactive controller as well. From these plots between different scenarios, it can be concluded that participants placed high importance on coordination (arriving at the same time) as opposed to simply speed or reactiveness. Between all the learning and validation trials, participants consistently rated ”slow” controllers (the robot arriving late) negatively.

VIII. CONCLUSIONS

A combination of the DMP framework for online trajectory generation and BO-based PL from direct user feedback allows for adaptive, responsive, and user-tailored robotic handovers. Following the results from Sect. IV and Sect. VI, batching relative user preferences with a refreshing scale can be highly

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

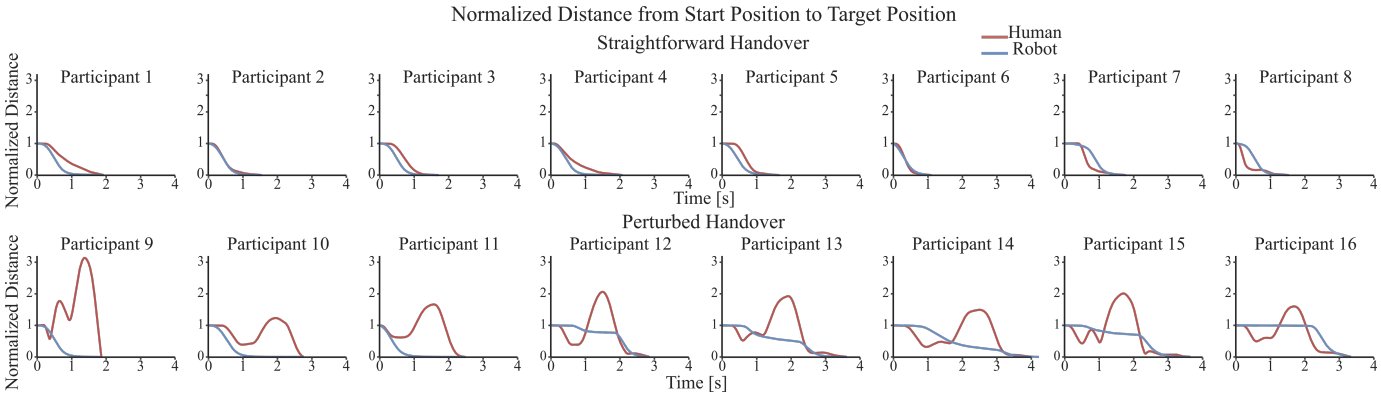

Fig. 4. Normalized (to distance between their respective Start positions and Handover position) human and robot distance over time in two scenarios. Human trajectories are in red, robot trajectories are in blue.

beneficial in short interactions. From the qualitative analysis of preferred robot behaviors, it can be apparent that users deem temporal coordination (partners arriving at the same time) as the key factor as opposed to other factors such as speed or reactiveness.

Practical robotic implementations would benefit from a combination of the proposed method with high-level FSM controllers (for example querying preference on interactions when some specific perturbation is detected). Furthermore, PL data could be used across different interactions with different participants to construct better priors, which might greatly benefit the BO approach.

REFERENCES

[1] M. Controzzi, H. Singh, F. Cini, T. Cecchini, A. Wing, and C. Cipriani, "Humans adjust their grip force when passing an object according to the observed speed of the partner's reaching out movement," *Exp Brain Res*, vol. 236, no. 12, pp. 3363–3377, Dec. 2018.

[2] A. Koene, A. Remazeilles, M. Prada, A. Garzo, M. Puerto, S. Endo, and A. Wing, "Relative importance of spatial and temporal precision for user satisfaction in Human-Robot object handover Interactions," in *AISB 2014 - 50th Annual Convention of the AISB*, Apr. 2014.

[3] A. Ijspeert, J. Nakanishi, and S. Schaal, "Movement imitation with nonlinear dynamical systems in humanoid robots," in *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No.02CH37292)*, vol. 2, May 2002, pp. 1398–1403 vol.2.

[4] F. Iori, G. Perovic, F. Cini, A. Mazzeo, E. Falotico, and M. Controzzi, "DMP-Based Reactive Robot-to-Human Handover in Perturbed Scenarios," *Int J of Soc Robotics*, vol. 15, no. 2, pp. 233–248, Feb. 2023.

[5] V. Ortenzi, A. Cosgun, T. Pardi, W. P. Chan, E. Croft, and D. Kulić, "Object Handovers: A Review for Robotics," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1855–1873, Dec. 2021, iEEE Transactions on Robotics.

[6] M. Prada, A. Remazeilles, A. Koene, and S. Endo, "Implementation and experimental validation of Dynamic Movement Primitives for object handover," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Sep. 2014, pp. 2146–2153, iSSN: 2153-0866.

[7] M. Wu, B. Taetz, Y. He, G. Bleser, and S. Liu, "An adaptive learning and control framework based on dynamic movement primitives with application to human–robot handovers," *Robotics and Autonomous Systems*, vol. 148, p. 103935, Feb. 2022.

[8] G. J. Maeda, G. Neumann, M. Ewerton, R. Lioutikov, O. Kroemer, and J. Peters, "Probabilistic movement primitives for coordination of multiple human–robot collaborative tasks," *Auton Robot*, vol. 41, no. 3, pp. 593–612, Mar. 2017.

[9] H. Ben Amor, G. Neumann, S. Kamthe, O. Kroemer, and J. Peters, "Interaction primitives for human-robot cooperation tasks," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, May 2014, pp. 2831–2837, iSSN: 1050-4729.

[10] C.-M. Huang, M. Cakmak, and B. Mutlu, "Adaptive Coordination Strategies for Human-Robot Handovers," in *Robotics: Science and Systems*, 2015.

[11] A. Kupcsik, D. Hsu, and W. S. Lee, "Learning Dynamic Robot-to-Human Object Handover from Human Feedback," in *Robotics Research: Volume 1*, ser. Springer Proceedings in Advanced Robotics, A. Bicchi and W. Burgard, Eds. Cham: Springer International Publishing, 2018, pp. 161–176.

[12] W. Chu and Z. Ghahramani, "Preference learning with Gaussian processes," in *Proceedings of the 22nd international conference on Machine learning - ICML '05*. Bonn, Germany: ACM Press, 2005, pp. 137–144.

[13] E. Brochu, V. M. Cora, and N. de Freitas, "A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning," Dec. 2010, arXiv:1012.2599 [cs].

[14] M. Tucker, M. Cheng, E. Novoseller, R. Cheng, Y. Yue, J. W. Burdick, and A. D. Ames, "Human Preference-Based Learning for High-dimensional Optimization of Exoskeleton Walking Gaits," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2020, pp. 3423–3430, iSSN: 2153-0866.

[15] M. Tucker, N. Csomay-Shanklin, W.-L. Ma, and A. D. Ames, "Preference-Based Learning for User-Guided HZD Gait Generation on Bipedal Walking Robots," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, May 2021, pp. 2804–2810, iSSN: 2577-087X.

[16] C. Wirth, R. Akrour, G. Neumann, and J. Fürnkranz, "A survey of preference-based reinforcement learning methods," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 4945–4990, Jan. 2017.

[17] H. J. Jeon, S. Milli, and A. Dragan, "Reward-rational (implicit) choice: A unifying formalism for reward learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 4415–4426, 2020.

[18] A.-L. Vollmer and N. J. Hemion, "A User Study on Robot Skill Learning Without a Cost Function: Optimization of Dynamic Movement Primitives via Naive User Feedback," *Frontiers in Robotics and AI*, vol. 5, 2018.

[19] B. Letham, B. Karrer, G. Ottoni, and E. Bakshy, "Constrained Bayesian Optimization with Noisy Experiments," *Bayesian Analysis*, vol. 14, no. 2, pp. 495–519, Jun. 2019, publisher: International Society for Bayesian Analysis.

[20] F. Iori, G. Perovic, F. Cini, A. Mazzeo, M. Controzzi, and E. Falotico, "DMP Based Perturbed Handover with Preferential Learning," *IEEE ICRA: Workshop on Machine Learning for Motion Planning*, May 2021.

[21] Z. J. Lin, R. Astudillo, P. Frazier, and E. Bakshy, "Preference Exploration for Efficient Bayesian Optimization with Multiple Outcomes," in *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. PMLR, May 2022, pp. 4235–4258, iSSN: 2640-3498.

[22] M. Balandat, B. Karrer, D. Jiang, S. Daulton, B. Letham, A. G. Wilson, and E. Bakshy, "BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 21 524–21 538.

[23] C. Cipriani, M. Controzzi, and M. C. Carrozza, "The SmartHand transradial prosthesis," *Journal of NeuroEngineering and Rehabilitation*, vol. 8, no. 1, p. 29, May 2011.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.